Simulating Correlated Random Variables

Philip M. Lurie and Matthew S. Goldberg
Institute for Defense Analyses
32nd DODCAS
2-5 February 1999

Outline

- Importance of correlations in cost analysis
- Logical consistency of correlation matrix
  - conditions for consistency
  - adjusting an inconsistent correlation matrix
- Use of simulation to account for correlations
- Methods for simulating correlated cost elements
- Lurie and Goldberg method
  - theory behind algorithm
  - numerical examples

Cost Elements are Often Correlated

- Spillover effects from one subsystem to another
  - induce positive correlations among cost elements
  - e.g., increase in airframe weight may require higher-thrust engines
- Schedule delays that necessitate paying overtime wage rates induce positive correlation between manhours and hourly wage rate
- Fungible costs that may be paid from one of several accounts
  - induce negative correlations across program phases
  - e.g., spare parts may be purchased using either Procurement or Operations and Maintenance (O&M) funds

Two General Methods to Account for Correlations

- Analytical methods
  - aggregate moments of underlying cost-element distributions
  - use mathematical analysis to estimate the distribution of total cost
- Simulation methods
  - generate random draws from each cost-element distribution
  - use empirical methods to estimate the distribution of total cost

When Analytical Methods are Practical

- Modelling only sum of components, e.g., total cost in a Work Breakdown Structure (WBS)
- Total cost obtained only by addition of random variables
  - rules out products, e.g., $Cost_i = Price_i \times Quantity_i$
- Under above conditions, a practical analytical method for estimating distribution of total cost is:
  - compute lower and upper bounds, mean, variance of sum from those of components
  - fit beta distribution to sum (this distribution is flexible enough to provide a good fit in most instances)
  - read percentiles of sum from fitted distribution

When Simulation Methods are Necessary

- Difficult to compute mean and variance of cost components
  - e.g., $Cost_i = Price_i \times Quantity_i$ where price and quantity may be correlated
- Cost components are not additive
  - e.g., completion times through a stochastic schedule network
  - nodal logic may depend on minimum or maximum times, not sum
  - difficult to compute mean and variance of total project cost and duration

Definitions

- Variance: $\sigma_i^2 = \sigma_{ii} = E(X_i - \mu_i)^2 = E(X_i^2) - \mu_i^2$
- Covariance: $\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j) = E(X_i X_j) - \mu_i \mu_j$
- Correlation: $\rho_{ij} = \sigma_{ij} / (\sigma_i \sigma_j)$
- Covariance matrix:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{pmatrix}$$

- Correlation matrix:

$$\begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & \rho_{kk} \end{pmatrix}$$

Logical Consistency of Correlation Matrix

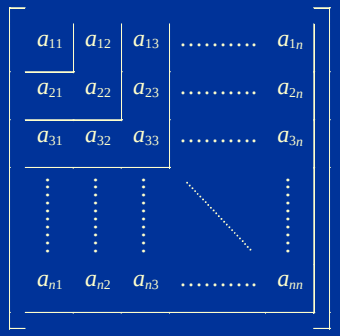
- When more than two cost elements are involved, there are constraints on possible values of correlations among them
- Example: three standardized (i.e., unit variance) cost elements *A*, *B*, *C* with *Corr*(*A*,*B*) = *Corr*(*B*,*C*) = 0.9
  - lowest possible correlation between *A* and *C* is 0.62
  - if correlation below 0.62 is specified, linear combination of costs with negative "variance" can be found
    - when $r = 0.5$, the quantity $D = .449A - .772B + .449C$ has "variance" of -0.047
- Inconsistencies can arise when
  - correlations derived from (multiple) expert opinion
  - not all correlations estimated from common data set

Positive-Definite Matrices

- What we are calling a logically consistent matrix is known in the mathematics literature as a positive (semi-)definite matrix
- A matrix $\boldsymbol{\Sigma}$ is said to be positive definite if $\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma} \mathbf{a} > 0$ for every vector $\mathbf{a} \neq \mathbf{0}$
  - if $\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma} \mathbf{a} \geq 0$ for all $\mathbf{a}$, $\boldsymbol{\Sigma}$ is said to be positive semi-definite
- If a vector random variable $\boldsymbol{X}$ has covariance matrix $\boldsymbol{\Sigma}$, then $Y = \mathbf{a}\boldsymbol{X}$ has covariance matrix $\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma} \mathbf{a}$
  - if $\boldsymbol{\Sigma}$ is inconsistent, a linear combination of the variables in $\boldsymbol{X}$ can be found with negative "variance," i.e., $\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma} \mathbf{a} < 0$

Checking for Logical Consistency

- The eigenvalues of Σ must all be greater than or equal to 0
- An equivalent condition for positive-definite matrices is that the principal minors of Σ must all be greater than 0
  - the kth principal minor is the determinant of the upper left $k \times k$ submatrix, $k = 1,2,\ldots,n$:



- If one or more of the principal minors is equal to 0, the matrix may or may not be positive semi-definite
  - the conditions for positive semi-definiteness are much more complex
- If one or more principal minors < 0, the matrix is inconsistent

Adjusting an Inconsistent Correlation Matrix to Make it Consistent

- Suppose a cost analyst proposes a correlation matrix that turns out to be inconsistent
  - it should not be used in estimating
  - the analyst probably isn't going to insist that his or her correlations are exact anyway
- To help the analyst out at this point, we have developed a method for adjusting a user-supplied "correlation" matrix that is inconsistent
- Resulting matrix is guaranteed to be
  - positive semi-definite (i.e., consistent)
  - as "close" as possible to user-supplied matrix

Adjusting the Correlation Matrix: Example 1

- Analyst's original correlation matrix:
  - determinant = –0.06

$$\begin{pmatrix} 1 & .9 & .5 \\ .9 & 1 & .9 \\ .5 & .9 & 1 \end{pmatrix}$$

- Analyst-supplied weighting matrix:
  - same degree of confidence in all correlations

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

- Adjusted correlation matrix:
  - eigenvalues = 2.52, 0.48, 0

$$\begin{pmatrix} 1 & .871 & .517 \\ .871 & 1 & .871 \\ .517 & .871 & 1 \end{pmatrix}$$

Adjusting the Correlation Matrix: Example 2

- *Corr*(*A*,*B*) = *Corr*(*B*,*C*) = 0.9, known with certainty
- *Corr*(*A*,*C*) = 0.5, very uncertain
- Same original correlation matrix: $\begin{pmatrix} 1 & .9 & .5 \\ .9 & 1 & .9 \\ .5 & .9 & 1 \end{pmatrix}$
- New weighting matrix: $\begin{pmatrix} 1 & 1 & .001 \\ 1 & 1 & 1 \\ .001 & 1 & 1 \end{pmatrix}$
- Adjusted correlation matrix: $\begin{pmatrix} 1 & .900 & .619 \\ .900 & 1 & .900 \\ .619 & .900 & 1 \end{pmatrix}$
  - eigenvalues = 2.62, 0.38, 0
  - uncertain elements bear full burden of adjustment

Methods for Simulating Correlated Cost Elements

- Functional relationships
  - Coleman and Gupta, TASC
  - specified by linking spreadsheet cells via formulas, e.g., production cost equals a (random) percentage of R&D cost of like item
  - unless ample historical data are available to estimate multivariate relationships, only simple pairwise relationships are likely to be specified
- Rank correlations
  - Iman and Conover, Sandia National Laboratories, *Communications on Statistics, Simulation and Computation*, Vol. 11, 1982
  - reorders independently generated random variables to achieve desired rank (Spearman) correlations
  - implemented in @RISK and Crystal Ball

Methods for Simulating Correlated Cost Elements (Continued)

- Completely specified distributions
  - Johnson, *Multivariate Statistical Simulation*, Wiley, 1987
  - complete multivariate structure must be specified; marginal distributions and correlation matrix may not be enough
  - correlations may already be determined given specification of marginals
    - e.g., Dirichlet distributions (multivariate generalization of beta)
- Partially specified distributions
  - Lurie and Goldberg, *Management Science*, Vol. 44, No. 2, 1998
  - useful for partially-specified distributions (only marginals and correlations need be specified)
  - uses Pearson (i.e., product-moment) correlation matrix

Simulating Multivariate Normal Distributions

- Statisticians have long known how to simulate multivariate normal distributions using Cholesky decomposition
- First generate a *k*-dimensional vector of independent normal random variables $\boldsymbol{X} \sim N(\mathbf{0},\mathbf{I})$
- If $\boldsymbol{X} \sim N(\mathbf{0},\mathbf{I})$, then $\boldsymbol{Y}=\mathbf{L}\boldsymbol{X} \sim N(\mathbf{0},\mathbf{L}\mathbf{L}^{\mathsf{T}})$
- If we want to generate $\boldsymbol{Y} \sim N(\mathbf{0},\boldsymbol{\Sigma})$, then need to find $\mathbf{L}$ such that $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$
  - Cholesky decomposition is a simple method for finding a lower-triangular matrix $\mathbf{L}$ such that $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$
  - once L has been determined, simply compute $\boldsymbol{Y}=\mathbf{L}\boldsymbol{X}$

Limitations of Cholesky Decomposition Method

- Method is valid for normal random variables only
  - if $\boldsymbol{X}$ is multivariate normally distributed, then linear combinations $\boldsymbol{Y}$ will also be normally distributed
- If method is misapplied to non-normal random variables
  - user-supplied means, variances, correlations are preserved
  - however, other distributional properties (e.g., modes, bounds, percentiles) are not preserved
  - if $X_1, X_2, \ldots, X_k$ have the "desired" (e.g., beta, triangular) distributions, the linear combinations $Y_1, Y_2, \ldots, Y_k$ will *not* inherit these distributions
- Method was at one time advocated by Aerospace Corp. (Book & Young, 24th DODCAS, 1990) but has since been disavowed

Lurie and Goldberg Method

- Adaptation of method originally proposed by Li and Hammond (*IEEE*, 1975)
- Designed to preserve user-specified marginal distributions and correlations among cost elements
  - all bounds, moments, and percentiles preserved
- Simulations can be done in an Excel spreadsheet, using Solver
- In some cases, it may be more practical to use a C++ or FORTRAN program with a non-linear optimization routine
  - large number of cost elements or Monte Carlo replications
  - input cost distributions without closed-form inverses

Initial Steps to Run the Lurie and Goldberg Method

- Generate *n* independent draws from a standard normal distribution for each variable: $\mathbf{X} \sim N(\mathbf{0},\mathbf{I})$
- Use Cholesky decomposition to transform independent normals into multivariate normals with user-supplied correlations: $\mathbf{Y} = \mathbf{LX} \sim N(\mathbf{0},\mathbf{LL}^{\mathsf{T}}) = N(\mathbf{0},\mathbf{R})$
  - constrain diagonal elements of $\mathbf{LL}^{\mathsf{T}}$ to equal 1
  - a positive semi-definite symmetric matrix with 1's along the diagonal is a correlation matrix
  - a good initial choice for **L** is the Cholesky factor of desired correlation matrix, $\mathbf{LL}^{\mathsf{T}}=\mathbf{R}$

Next Steps to Run the Lurie and Goldberg Method

- Apply standard normal distribution function Φ to each marginal normal distribution: $U_i = \Phi(Y_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Y_i} e^{-\frac{1}{2}y^2} dy$
  - results in correlated uniform random variables
  - correlations will be different from those of original variables
- Invert each uniform distribution using the user-specified marginal distributions F_i: $V_i = F_i^{-1}(U_i)$
  - results in correlated random variables with user-specified marginal distributions
  - original correlations will be further distorted

Final Steps to Run the Lurie and Goldberg Method

- Compute correlations ($\hat{r}_{ij}$) among final transformed variables (V_i, V_j)
- Compute a distance measure between user-supplied correlation matrix and transformed correlation matrix computed in the previous step:

$$D = \sqrt{[2/\{k(k-1)\}]\sum_{i=2}^{k}\sum_{j=1}^{i-1}(\hat{r}_{ij} - r_{ij})^2}$$

- Iterate over elements of Cholesky factorization matrix (**L**) to minimize the above distance measure
  - constrain diagonal elements of $\mathbf{LL}^{\mathsf{T}}$ to equal 1
  - requires a non-linear optimization routine such as Microsoft Excel Solver or a specially-written FORTRAN or C++ program

Concise Summary of Lurie and Goldberg Method

- Find matrix **L** such that series of transformations

$$\underset{indep.normal}{X} \xrightarrow{\mathbf{L}} \underset{mult.normal}{Y} \xrightarrow{\Phi} \underset{uniform}{U} \xrightarrow{F^{-1}} \underset{desired}{V}$$

  lead to random variables with desired correlations and marginal distributions
  - **L**: Cholesky factor transforms independent normals to correlated normals
  - Φ: normal c.d.f. transforms correlated normals to correlated uniforms
  - F^{-1}: transforms correlated uniforms to correlated random variables with desired marginal distributions F
- Because Φ and F^{-1} are non-linear transformations, the correlations among the V's will differ from the correlations among the Y's
- Iterate over **L** to achieve desired correlations among V's

Theoretical Convergence Guarantees

- Distance measure (between user-supplied and transformed correlation matrices) is bounded below by zero and possesses a minimum
  - minimum distance is zero if user-supplied correlations can be theoretically achieved by transforming correlated normals
  - minimum distance may be strictly positive
- Algorithm is guaranteed to converge to global minimum if:
  - distribution functions are all continuous
  - Gauss-Newton or any standard quasi-Newton method is used to minimize the distance measure
  - starting values are sufficiently close to the minimum
    - Cholesky factor of desired correlation matrix, $\mathbf{L}\mathbf{L}^{\mathsf{T}}=\mathbf{R}$, is usually close enough

Performance of Lurie and Goldberg Method

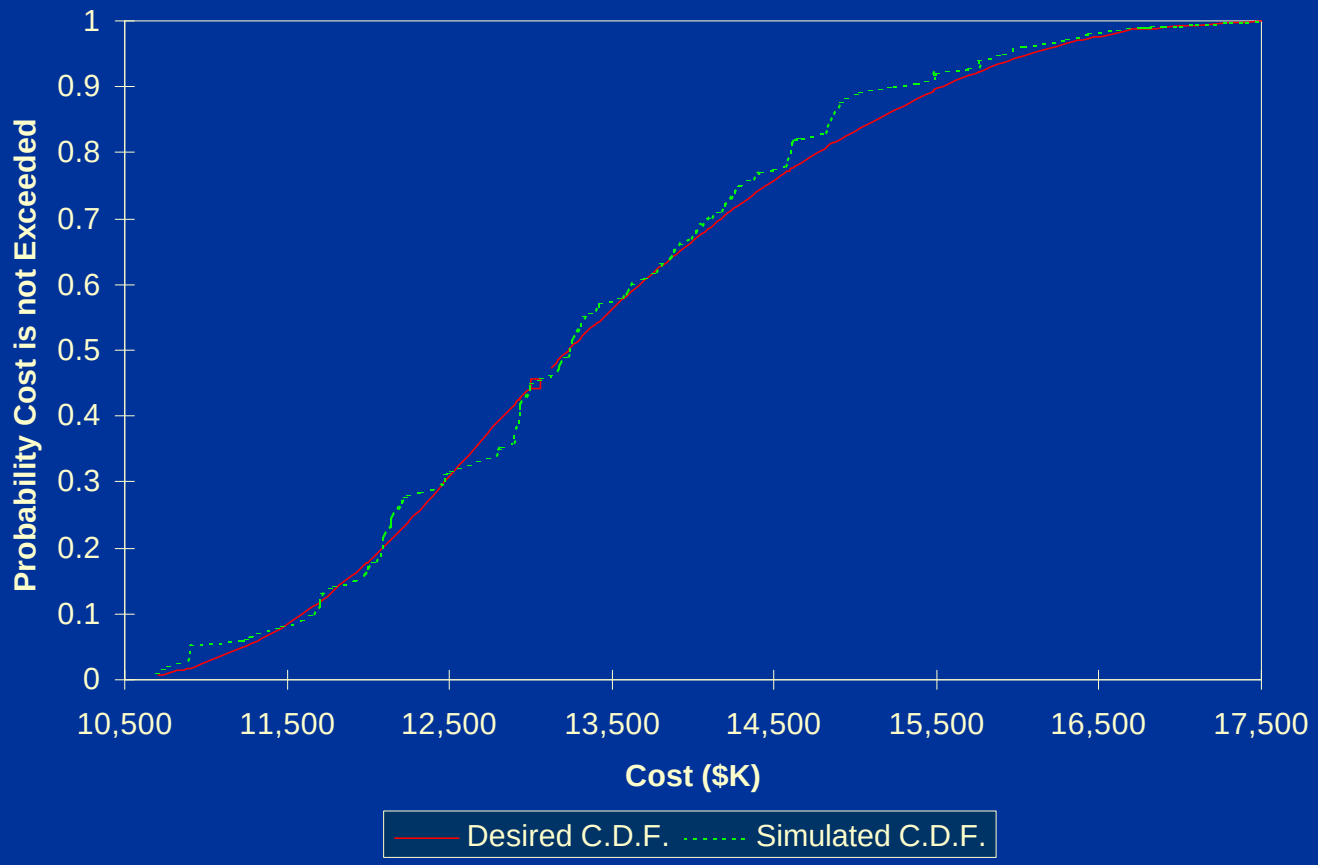
- Accuracy:
  - preserves user-specified distributions and correlations, particularly for large sample sizes
  - validated on several multi-dimensional test problems
  - performance does not degenerate with increasing number of cost elements
- Speed:
  - depends on dimension of problem and CPU speed
  - faster if user-supplied distributions have closed-form inverse (e.g., triangular) rather than requiring numerical approximation (e.g., beta)

Example: First-Unit Cost WBS for 600-lb. UHF Satellite

All distributions assumed triangular, with the following parameters:

	Cost ($K)		
Cost Element	Lower Bound	Mode	Upper Bound
Attitude Control	1,676	1,942	2,453
Electrical Power Supply	3,469	4,329	5,287
Telemetry, Tracking and Command	860	1,014	1,671
Structure and Thermal	366	596	963
Apogee Kick Motor	201	314	402
Digital Electronics	5,433	8,431	8,828
Communications Payload	2,228	2,425	3,713
Integration and Assembly	544	691	1,011
Program Support	10,410	12,428	17,400
Launch Operations and Orbital Support	639	914	1,030

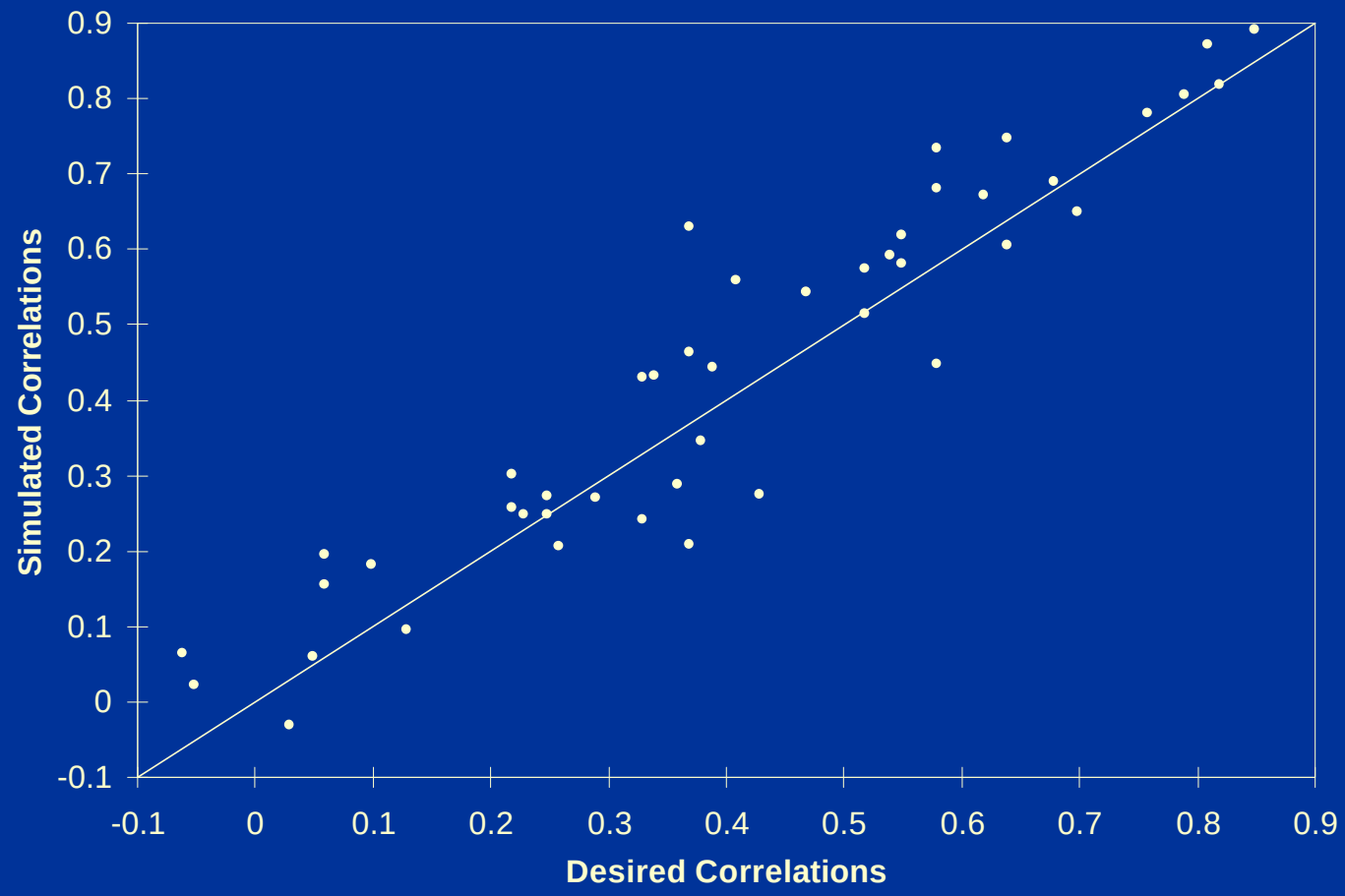
Simulated vs. Desired Distribution of Program Support Cost
(Sample Size n=100)



Note: Program Support is worst-fitting among all ten cost elements

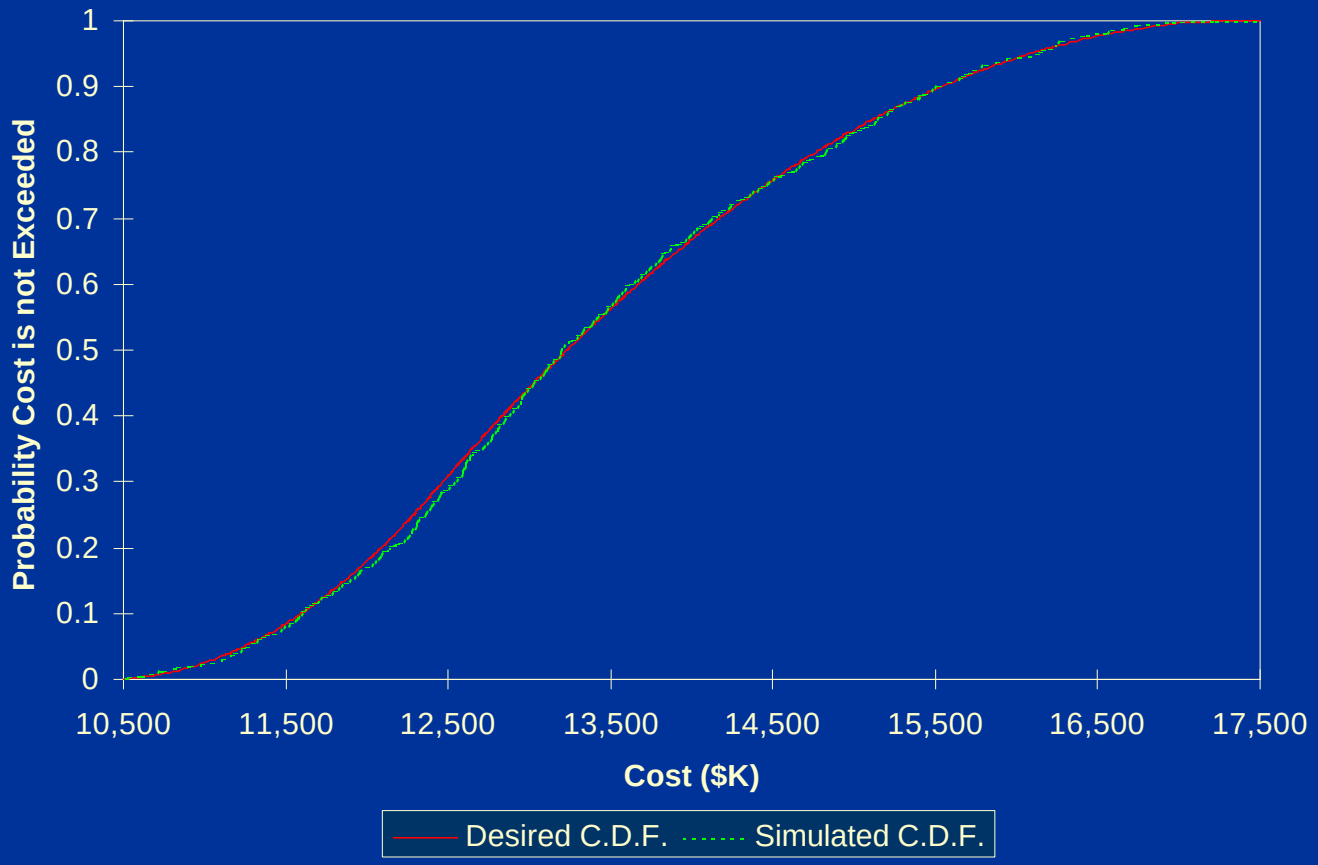
Simulated vs. Desired Correlations

(Sample Size n=100)



Simulated vs. Desired Distribution of Program Support Cost (Sample Size n=1,000)

- Correlation matrix reproduced “exactly”



Note: Program Support is worst-fitting among all ten cost elements

Implementing Lurie-Goldberg

- Lurie-Goldberg can be implemented in an Excel spreadsheet or in a higher-level programming language (such as FORTRAN or C++)
  - no user-friendly software currently available
    - in-house version, which simulates multivariate triangular distributions, contains about 250 lines of FORTRAN code including calls to commercially-available factorization and optimization routines
  - needs an interface to allow users to choose from several distributions
- Most practical environment would be as an add-in to Excel or other spreadsheet package
  - @RISK and Crystal Ball developers might be persuaded to include it if there were sufficient interest

Simulation Software Vendors

@RISK:

Palisade Corporation
31 Decker Road
Newfield, New York 14867
1-800-432-7475
Fax: 607-277-8001
www.palisade.com

Crystal Ball:

Decisioneering Inc.
1515 Arapahoe Street, Suite 1311
Denver, Colorado 80202
1-800-289-2550
Fax: 303-534-4818
www.decisioneering.com

References

- Book and Young, The Aerospace Corporation, 24th DODCAS, 1990
- Coleman and Gupta, TASC, 28th DODCAS, 1994
- Iman and Conover, Sandia National Laboratories, *Communications on Statistics, Simulation and Computation*, Vol. 11, 1982
- Johnson, *Multivariate Statistical Simulation*, Wiley, 1987
- Li and Hammond, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 5, 1975
- Lurie and Goldberg, *Management Science*, Vol. 44, No. 2, 1998